HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Leading Indicator Search for Time Series Forecasting

Lucas Möller

Hochschule Darmstadt - Fachbereiche Mathematik und Naturwissenschaften & Informatik

Motivation & Research Questions

Accurate forecasts of demand play an important role in many businesses and industries. Especially in the retail sector, these forecasts build the basis for planning various supply-chain activities such as stock management or the allocation of scarce resources and personnel [1]. In this context, the underlying data are often time series. What makes time series data special is that successive observations are usually not independent. In order to make predictions of the future course of time series, established time series models therefore attempt to model the inherent structures and patterns based on past observations.

Here, one important factor, namely information from external data sources, is often neglected. However, the inclusion of external data sources and the identification of indicators, that have a leading effect on the time series being forecast, can improve the sales forecast accuracy. If these indicators exhibit similarities in terms of patterns and temporal structures, they may contain leading context information that explain some of the historical variation [2]. Selecting appropriate leading indicators and their respective lead order, however, is not trivial. For this reason, the purpose of this work was to examine an approach, in which external time series can be merged and processed so that they potentially lead to an improving forecast performance. The motivation and purpose led to the two core research questions of this work:

1. Does the integration of external data sources and leading indicators contribute to improving the accuracy of forecasts?
2. When is it beneficial to add external regressors and which conditions are critical for success?

Methodology

The research questions were answered in the context of an experiment. Based on the cross-correlation as a similarity measure, leading indicators from an external online open data source were determined for the time series of a German retail company. These leading indicators were then incorporated individually as external regressors into a *ARIMA* time series model, converting a univariate model to a bivariate *ARIMAX* model. Based on the achieved forecast errors, the *Mean Absolute Error* (MAE) metric was calculated to compare the model accuracies for each time series. The comparison of the forecasting performance was intended to provide information on which factors are responsible for the successful inclusion of leading indicators. Retail sales, in particular, often exhibit strong seasonal variations, consequently making an effective modeling of retail sales time series a challenging task [3]. For this reason, in parallel with the calculation of the cross-correlations, the seasonal strengths of all time series were also computed according to the following formula:

$$F_S = max\left(0, 1 - \frac{Var(R_t)}{Var(S_t + R_t)}\right),$$

where $Var(\cdot)$ is the variance of the corresponding time series component. A value of F_S close to 0 indicates almost no seasonality within the time series. Contrarily, a series with strong seasonality will have F_S close to 1 [4]. The goal was to investigate the influence of seasonal components on the outcome of the experiment. The obtained results were validated with the three multivariate analysis methods *Multiple Linear Regression*, *Analysis of Variance* (ANOVA) and *Align Rank Transform Contrasts* [5].

Datasets

The experiment was based on two datasets. One dataset was provided by a German retail company which is active in the discounter and construction market as well as in the consumer electronics business. The retail sales time series cover a period of roughly three years. After consultation with the company, only products of the superordinate retail area *Fast Moving Consumer Goods* (FMCG) were considered in the experiment. FMCG are relatively low-priced products with a high turnover, which satisfy immediate wants and needs. In addition, a single store was selected that had the most FMCG products sold. Ultimately, there were 3335 unique products that lead to 3335 sales time series in total.

The second dataset was retrieved from *Google Trends* as an online open data source. Google Trends is a tool developed by *Google LLC* that provides fine-grained data on the popularity of customer queries on certain search terms in the Google search engine [6]. The search terms used to create the Google Trends dataset were collected based on *unigrams* of the descriptions of the retail sales products. The premise was, if customers use the Google search engine to search for certain FMCG products, it may be an indication that they have an immediate need and may purchase these products in the near future. Since the product descriptions from the retail sales dataset revealed serious differences in terms of quality, some preprocessing steps (e.g., removal of special characters and quantity information, filtering of English and German stopwords) were necessary in order to obtain the unigrams. Table 1 shows three exemplary product descriptions and their resulting unigrams.

Product descriptions	*Unigrams*
Knorr Rahmsauce Braten & Schmoren 250ml	[knorr, rahmsauce, braten, schmoren]
Bratkartoffeln 400g	[bratkartoffeln]
Bio Lacroix Paste 40g, Gemüse	[lacroix, paste, gemüse]

Table 1: Product descriptions and corresponding unigrams

There were 3335 unique products with corresponding product descriptions in the retail sales dataset. After applying the preprocessing steps, 9274 unigrams in total could be extracted. These unigrams served as keywords to generate the Google Trends dataset. Here, the Python library pytrends was utilized as an Application Programming Interface to request the keyword related time series within a Python program. However, not every keyword leads to search queries in Google and thus to corresponding time series data. From 9274 unigrams in total only 1848 keywords had search queries in Google. Therefore, the Google Trends dataset eventually consisted of 1848 keyword related time series.

After the Google Trends dataset was created, the cross-correlation between each retail sales time series and every keyword time series was computed for a number of lags. The keyword that had the highest cross-correlation with the respective retail sales time series was set to be the leading indicator for this series. Eventually, the similarity matching and the leading indicator search can be summarized by the following pipeline illustrated in figure 1.

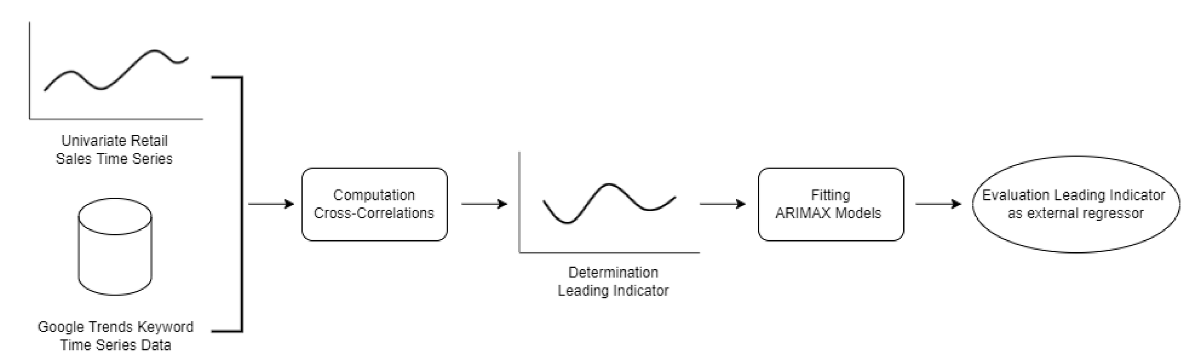
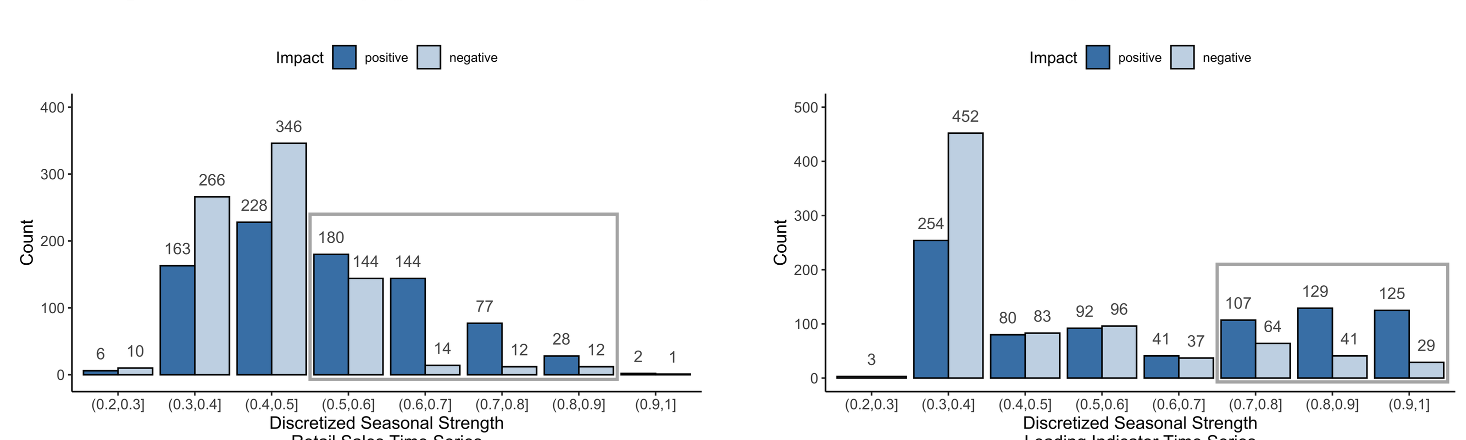


Figure 1: Pipeline of key experimental steps.

Results

The retail sales dataset included time series of products that were sold only irregularly (the proportion of the value 0 was high). To ensure that such time series did not distort the results of the experiment, they were excluded from the dataset beforehand. Therefore, retail sales time series, in which the value 0 had a relative proportion of $\geq$ **50%** over the entire period and $\geq$ **25%** in year 2021 were eliminated. The remaining time series were considered efficient. Ultimately, there were 1633 retail sales time series for which results were obtained in the experiment.

The comparison of the ARIMA and ARIMAX models has shown that the inclusion of the leading indicators has led to an improvement of the forecast accuracy in 51% and to a deterioration in 49% of the cases. Therefore, the first question if the inclusion of external time series as leading indicators can improve the forecast accuracy can be answered with yes. However, there is no clear tendency and the improvements can be solely by chance. In order to exclude this chance, individual time series with both noticeable improvements and significant deteriorations have been examined. The examination has shown that the outcome of the inclusion possibly depends on the seasonal strengths of the retail sales and its corresponding leading indicator time series. Moreover, the cross-correlation achieved in the similarity matching seems to play a subordinate role. This analysis was then extended to all time series. It was found that positive and negative results were obtained in all cross-correlation ranges. There were no ranges where the proportion of the positive impact ($MAE_{ARIMAX} < MAE_{ARIMA}$) excels and vice versa. The ratio between positive and negative impact was balanced. For the seasonal strength, however, the proportion of the positive impact was notably greater in higher ranges as demonstrated in figure 2.



(a) Retail Sales Time Series. (b) Leading Indicator Time Series.

Figure 2: Discretized seasonal strengths separated by impact of leading indicator.

Based on these exploratory results, the two following hypotheses were formulated, which were tested with the three multiple analysis methods:

- There is no relationship between cross-correlation and the outcome of the inclusion. Improvements and deteriorations are both obtained equally at low and high cross-correlation values.
- If both the retail sales and leading indicator time series exhibit strong seasonal patterns, then the probability of a positive impact increases. In such cases, the inclusion of the external regressor can be particularly beneficial.

All three analysis methods concluded that when the seasonal strengths of the retail sales time series and their leading indicator series are high, the probability of achieving an improving forecast accuracy increases. Especially the two-way interaction between the dichotomized seasonal strength variables appeared to be statistically significant as illustrated in figure 3.

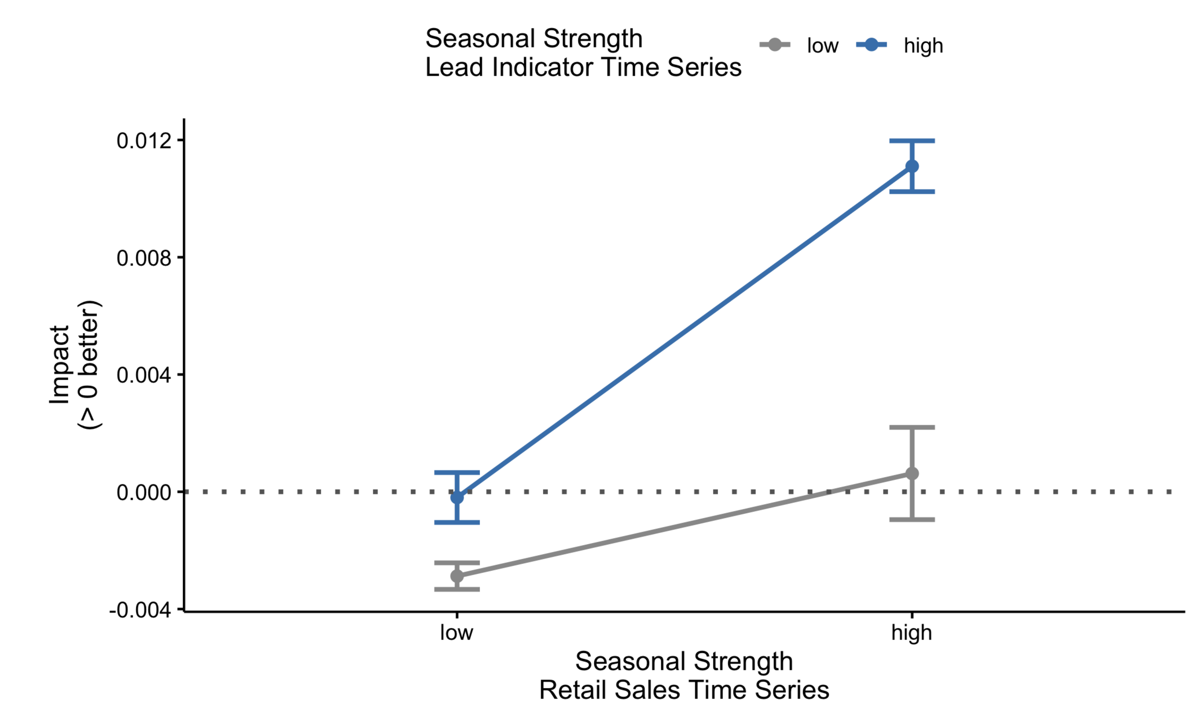


Figure 3: Two-way interaction of dichotomized seasonal strength variables.

It turned out that the seasonal pattern is a critical success factor for the experiment. On the other hand, the evaluations of the multiple analysis methods have demonstrated that the cross-correlation exerts no statistically significant influence. Improvements and deteriorations were both obtained equally at low and high cross-correlation values.

However, it must be pointed out that the fundamental model assumptions of the parametric methods Multiple Linear Regression and ANOVA were violated. Therefore, the results and conclusions from these two methods are suspect and potentially invalid. The analysis results of the nonparametric alternative Align Rank Transform Contrasts, however, could confirm the results of the two parametric methods.

Ultimately, the hypotheses stated can be confirmed. There is no relationship between cross-correlation and the outcome of the integration. A higher similarity does not guarantee a higher forecast accuracy with ARIMAX models. Contrarily, the interaction between the seasonal strength variables is statistically significant in a way that higher values lead to positive effects. One possible interpretation is that the historical variation of highly seasonal retail sales time series can be explained with information provided by the strong seasonal components of their leading indicators. The short-term development of these leading indicators may anticipate upcoming seasonal patterns of the retail sales time series. As the results show, these information can be particularly valuable for forecast accuracy.

These results make it possible to answer the second question, when it is useful to include external regressors and which conditions are critical for success.

Conclusion

Based on the results obtained in this work, it can be concluded that high and low cross-correlations can contribute to an improving forecast accuracy, provided the merged time series have strong seasonal patterns. Time series with high seasonal strengths are therefore suitable candidates for leading indicator search. For this reason, it is important to consider the seasonal strength of time series within the matching process in further research questions. In addition, further similarity measures next to the cross-correlation should be tested.

References

[1] S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37–45, 2018.

[2] J. H. Stock and M. W. Watson, "Forecasting inflation," Working Paper 7023, National Bureau of Economic Research, March 1999.

[3] C.-W. Chu and G. P. Zhang, "A comparative study of linear and nonlinear models for aggregate retail sales forecasting," *International Journal of Production Economics*, vol. 86, no. 3, pp. 217–231, 2003.

[4] R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice.* Melbourne, Australia: OTexts, 3 ed., 2021.

[5] L. A. Elkin, M. Kay, J. J. Higgins, and J. O. Wobbrock, "An aligned rank transform procedure for multifactor contrast tests," in *The 34th Annual ACM Symposium on User Interface Software and Technology*, Association for Computing Machinery, 2021.

[6] E. Cebrián and J. Domenech, "Is google trends a quality data source?," *Applied Economics Letters*, pp. 1–5, 2022.